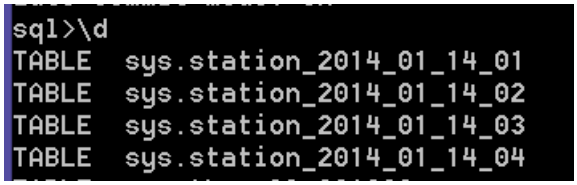
Three tasks are dealt with last week. The first one is to import much more data into the memory database MonetDB and test its performance. The Second is to analyze Twitter data for the student of Prof. Wei. The last one is to find and collect query patterns of mobility data.

It almost took me two days to import mobility data of one day (14GB) into MonetDB from scratch. I made a program to convert the binary file into an ASCII text file, which can be recognized by MonetDB. Because MonetDB only accepts files in specific forms, I have failed for two times before success. Four tables are created and each table contains data in six hours, as shown in Figure 1.



Then, some simple test were taken and returned good results, as follows.

Query	Performance
SELECT id FROM station_2014_01_14_01 LIMIT 400; (69679731 rows in total)	1.969 ms
SELECT * FROM station_2014_01_14_02 WHERE id = 460077059245077;	44.936 ms
SELECT * FROM station_2014_01_14_02 WHERE cell =16375;	96.44 ms

From the time we can find that it is better than Hive or existing DBMS. We will regard this memory database as a standard to measure our data management system in plan.

Sentiment analysis is very important in social media. Wei's group wants to know opinions on global warming. We hold twitter data of 5 months respectively in 2011 and 2012. After searching "global warming" in database, 400 + 1000 tweets return. Stanford Linguistics group publishes a sentiment tutorial, including many different methods to deal with text. Yet, most methods can not provide satisfactory answers. I have dispatched this work to a student to find appropriate methods.

I also collect some query patterns during the last week. On Monday, we talked about how to determine a person indoor or outdoor. We can define a pattern on temporal and spatial. By temporal, the pattern should last a long term. By spatial, the pattern should span a large area. In the Nature paper "Unique in the Crowd: The privacy bounds of human mobility", the author also mentions some query patterns, such as "increasing the size of a region, aggregating neighbouring cells into clusters of v cells, or by reducing the dataset's temporal resolution, increasing the length of the observation time window to h hours", and increasing the number of antennas used for localizations. The paper "Efficient Range Distribution Query for Visualizing Scientific data"reported by Ding also inspires me. It is very difficult to build a general data management system; however, we can construct a special system for an application. For large dataset, the key points are how to partition the dataset and execute query seperately. In conclusion, we can concern about patterns in three perspectives, i.e. time windows, spatial windows and spans of aggregations. Therefore, the query model can be built on these patterns.

In next week, I will create a demo of this query model to analyze some phenomenon interactively.